Morphosyntactic resources for automatic speech recognition

Stéphane Huet, Guillaume Gravier, Pascale Sébillot

Morphosyntactic resources for automatic speech recognition

Stéphane Huet, Guillaume Gravier, Pascale Sébillot

IRISA/Université de Rennes 1, IRISA/CNRS, IRISA/INSA de Rennes
Campus de Beaulieu, F-35042 Rennes Cedex, France
{shuet, ggravier, sebillot}@irisa.fr

Abstract

Texts generated by automatic speech recognition (ASR) systems have some specificities, related to the idiosyncrasies of oral productions or the principles of ASR systems, that make them more difficult to exploit than more conventional natural language written texts. This paper aims at studying the interest of morphosyntactic information as a useful resource for ASR. We show the ability of automatic methods to tag outputs of ASR systems, by obtaining a tag accuracy similar for automatic transcriptions to the 95-98 % usually reported for written texts, such as newspapers. We also demonstrate experimentally that tagging is useful to improve the quality of transcriptions by using morphosyntactic information in a post-processing stage of speech decoding. Indeed, we obtain a significant decrease of the word error rate with experiments done on French broadcast news from the ESTER corpus; we also notice an improvement of the sentence error rate and observe that a significant number of agreement errors are corrected.

1. Introduction

Automatic speech recognition (ASR) systems generate transcriptions of spoken productions, by converting a speech signal into a sequence of words. These tools offer the possibility to produce spoken texts from audio streams with a few human resources. However, despite some progress, the outputs generated by this automatic process still contain errors about words to recognize, which leads to noisy texts. In this article, we show that part of speech (POS) tagging, as a linguistic resource which is widely spread in the natural language processing (NLP) community, improves the quality of transcriptions.

Tagging aims at associating each word or locution of a sentence with a class that brings information about its part of speech and often about its morphology (number, gender, tense…). This process has been seldom studied for spoken corpora (Valli and Véronis, 1999), especially for texts produced by ASR systems; it has however several interests. First, many tasks do require POS tagging as a preliminary step to exploit audio streams: spoken document indexing, topic tracking, summarization… Moreover introducing new linguistic knowledge, such as morphosyntactic information, offers some prospects to obtain better transcriptions. A significant number of transcription errors could be corrected by information about gender and number agreement. In particular, in the French language, nouns, adjective, and verbs are very often inflected for number, gender, or tense into various homophone forms; this property increases the interest of POS to reduce the number of misrecognized words.

In spite of its relevance for transcriptions produced automatically, POS tagging of such noisy texts rises several difficulties. Oral output has characteristics, such as repetitions, revisions or fillers that make it not straightforward. Additional difficulties come from the fact that automatic transcriptions are not segmented into sentences. These texts also lack punctuation and, in the case of some ASR systems such as ours, capitalization. Besides, ASR generates transcriptions with misrecognized words, which lead to ungrammatical word sequences, whereas NLP techniques are usually applied to correct texts.

In this paper that aims at showing the interest of morphosyntactic resources for ASR, we first give a very brief overview of the ASR process (Section 2). Section 3 describes characteristic transcription errors, which can disturb the use of NLP techniques or which can be corrected through the use of morphosyntax. Then, Section 4 shows the ability of POS tagging to face the oral characteristics of transcription hypotheses. Finally, Section 5 quantitatively demonstrates that tagging is useful to improve the quality of transcriptions.

2. The basic principles of automatic speech recognition

Most automatic speech recognition systems rely on statistical models of speech and language to find out the best transcription, *i.e.*, word sequence, given a (representation of the) signal y, according to

$$\widehat{w} = \arg\max_{w} p(y|w)\ P[w]\ . \qquad (1)$$

Language models (LM), briefly described below, are used to get the prior probability $P[w]$ of a word sequence w. Acoustic models, typically continuous density hidden Markov models (HMM) representing phones, are used to compute the probability of the acoustic material for a given word sequence, $p(y|w)$. The relation between words and acoustic models of phone-like units is provided by a pronunciation dictionary which lists the words recognizable by the ASR system along with their corresponding pronunciations. Hence, ASR systems operate on a closed vocabulary whose typical size is between 60,000 and 100,000 words or tokens. Because of the limited size of the vocabulary, word normalization is often used to limit the number of out-of-vocabulary words, for example by ignoring the case or by splitting compound words. The consequence is that the vocabulary of an ASR system is not necessarily suited for natural language processing.

Although the number of hypotheses is limited to the words of this vocabulary, Equation (1) leads to consider a hypothesis space that dramatically increases with the number of

consecutive words to recognize. To solve this problem, this equation is solved for quite short utterances, by segmenting the speech stream into breath-groups, where the definition of this unit is based on the energy profile in order to detect breath intakes. Let us stress that this segmentation only uses an acoustic cue —silence duration— and is not based on syntactic and grammatical considerations, even through breath pauses and grammar are related.

As mentioned previously, the role of the language model is to define a probability distribution over the set of possible hypotheses according to the vocabulary of the system. As such, the language model is a key component for a better integration between ASR and NLP. ASR systems typically rely on N-gram based language models because of their simplicity which makes the maximization in (1) tractable. The N-gram model defines the probability of a sentence w_1^n as

$$P[w_1^n] = \prod_{i=1}^{n} P[w_i|w_{i-N+1}^{i-1}] \ , \qquad (2)$$

where the probabilities of the sequences of N words $P[w_i|w_{i-N+1}^{i-1}]$ are estimated from large text corpora. Because of the large size of the vocabulary, observing all the possible sequences of N words is impossible. A first approach to circumvent the problem is based on smoothing techniques, such as discounting and back-off, to avoid null probabilities for events unobserved in the training corpus. Another approach rely on N-gram models based on classes of words (Brown et al., 1992) where a N-gram model operates on a limited set of classes, and words belong to one or several classes. The probability of a word sequence is then given by

$$P[w_1^n] = \sum_{t_1 \in \mathcal{C}(w_1) \ldots t_n \in \mathcal{C}(w_n)} \prod_{i=1}^{n} P[w_i|t_i] P[t_i|t_{i-N+1}^{i-1}] \ , \qquad (3)$$

where $\mathcal{C}(w)$ denotes the set of possible classes for a word w.

In practice, (1) is evaluated in the log-domain and the LM probabilities are scaled in order to be comparable to acoustic likelihoods, thus resulting in the following maximization problem

$$\widehat{w} = \arg\max_{w} \ \log p(y|w) + \beta \ \log P[w] + \gamma \ |w| \ , \qquad (4)$$

where the LM scale factor β and the word insertion penalty γ are empirically set.

The ultimate output of an ASR system is obviously the transcription. However, transcription alternatives can also be obtained. This information might prove useful for NLP as it can help to avoid error-prone hard decisions from the ASR system. Rather than finding out the sole best word sequence maximizing (4), one can output a list of the $\mathcal{N}$-best word sequences thus keeping track of the alternative transcriptions that were discarded by the system. For a very large number of transcription hypotheses, these $\mathcal{N}$-best lists can be conveniently organized as word graphs where each arc corresponds to a word, or as confusion networks where the identical words of the hypotheses of the $\mathcal{N}$-best lists are aligned (Mangu et al., 2000).

Two measures are frequently used to evaluate the results of the output of ASR systems: the word error rate (WER) and the sentence error rate (SER). Both are computed from the alignment of a reference transcription made by a human annotator with the transcription generated by the ASR system. The WER is obtained by counting the number of insertion, deletion and substitution errors observed between these two outputs, while the SER gives the proportion of breath-groups that are transcribed without a single misrecognized word.

3. Typical transcription errors

This section is dedicated to the common recognition errors done by ASR systems, with the willing to show the proportion of misrecognized words *a priori* rectifiable by morphosyntactic resources. To illustrate the significance of each kind of errors, we study here a 30-minute excerpt of French radio broadcast news, transcribed by our ASR system with a WER of 17.8 %.

A first type of errors corresponds to numerous consecutive misrecognized words, which leads to ungrammatical breath-groups, frequently along with out-of-topic words. These errors are explained by two main sources. The first is bad acoustics, that makes some words particularly difficult to recognize: two extracts with a noisy background, which globally have a 2-minute duration, represents 15.3 % of the transcription errors for the whole corpus studied. The significance of these errors greatly varies for broadcast news with the proportion of interviews recorded outside radio studios and their reduction can only be made at the acoustic level. The second source of errors is the presence of named entities or technical terms. These words are often difficult to recognize, since their huge potential number prevents from having a good coverage with a vocabulary of a limited size, all the more that named entities appear and disappear across the time according to the news. The named entities and technical terms represent 11.3 % of the overall errors measured on the corpus; these errors would have been more numerous if the broadcast news studied here had been recorded in a period different from that of the corpus used to select the vocabulary of the ASR system.

The size of the breath-group has also an influence over the number of misrecognized words. Indeed, short breath-groups, which offer a more limited context to select the word hypotheses through the LM, tend to result in more transcription errors. On the 30-minute corpus, the breath-groups made of at most 5 words represent 4.4 % of the transcribed words, but 5.2 % of the errors. We show in Figure 1 how the WER varies according to the size of breath-groups to recognize with respect to the WER; this graph exhibits a significant increase of errors for breath-groups with less than 5 words. Correcting this kind of errors implies to re-segment the signal into larger units.

Some very problematic words to recognize are short grammatical words. They are often difficult to detect by the acoustic model due to their very fast pronunciation, which lets the language model alone to recognize them; the word insertion penalty also tries to introduce knowledge about their frequency in a given breath-group by penalizing or favoring the insertion of words, particularly short words.

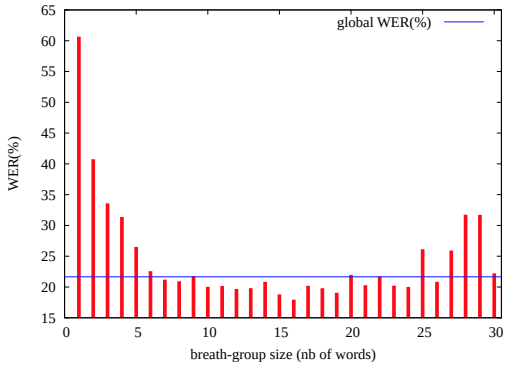

Figure 1: WER as a function of breath-group size on a 4-hour corpus including the 30-minute corpus studied here.

However, some uses of grammatical words are badly modeled by word-based LMs, as they require to understand the context to choose the correct forms. On the 30-minute corpus, we noticed that the most problematic grammatical words are the auxiliaries "*avoir*" ("to have") and "*être*" ("to be"), prepositions, conjunctions and determiners. Errors on short grammatical words are particularly numerous as they affect more than one out of five breath-groups in our corpus.

We also noticed other grammatical errors. Some are confusions about tense and mood for verbs; these errors represent 2.4 % of the overall errors observed on the 30-minute corpus. Besides, we observed errors explained by how speech is segmented to be decoded. For instance, the segmentation according to an acoustic criterion rather than a more linguistic one leads sometimes to recognize breath-groups that begin with verbs; this case tends to generate errors by using LMs that favor other classes of words at the beginnings of breath-groups. Furthermore, ungrammatical hypotheses frequently appear by misrecognizing words when repetitions or repairs occur in the speech stream. Errors due to these two kinds of disfluencies represent 2.6 % of those observed on the 30-minute corpus; this proportion would have been more important if the studied extract had contained more spontaneous speech.

Finally, a last significant class of errors is related to errors about gender or number agreement, and confusion about infinitive and past participle. This kind of transcription errors, as mentioned in the introduction, is particularly important for the French language, for which inflections tend to produce several homophone forms; it represents for the 30-minute corpus 11.7% of overall errors. Among this kind of errors, some of them require anaphora resolution (2.9 % of overall errors), with references across several breath-groups. Others are explained by agreements between two entities that belong to different breath-groups. We finally count 76 transcription errors, representing 6.5 % of the overall errors, rectifiable by considering independently each breath-group. 5 of them (0.4 %) are actually impossible to correct without expanding the vocabulary of ASR system since this lexicon does not contain the correct forms.

This description of typical errors exhibits the potential interest of morphosyntactic information. If some transcription errors require to know the context and are out of scope of this kind of knowledge, morphosyntax can bring syntactic constrains over the use of word classes, especially grammatical ones. It is also relevant to reduce agreement errors. On the 30-minute corpus analyzed, 6.1 % of misrecognized words are due to agreement errors and are rectifiable by considering independently each breath-group or without having to expanding the vocabulary of ASR system; correcting these errors would represent for this extract an absolute decrease of the WER of 1.1 %.

In the next parts of this paper, we demonstrate quantitatively that POS tagging can be performed reliably on the output of ASR systems in spite of the misrecognized words and that it can be used to reduce the number of transcription errors.

4. POS tagging of oral corpora

As oral output has specificities that are likely to disturb taggers, we first demonstrate that such noisy texts can be reliably tagged. To this end, we developed a morphosyntactic tagger specifically for the outputs of our ASR system. We evaluated it on the ESTER corpus that consists of French-speaking radio broadcast news (Galliano et al., 2005).

4.1. Method

We built a morphosyntactic tagger based on the popular HMM technique (Merialdo, 1994), where tagging is expressed as finding out, for each sentence, the most probable POS tag sequence, among all the possible sequences. Other tagging methods, such as maximum entropy models (Ratnaparkhi, 1996) or support vector models (Giménez and Màrquez, 2004), have been conceived since the first uses of HMM-based taggers. However, a study led by Brants (2000) shows that HMM gives results comparable to other techniques.

HMM-based tagging can be seen as choosing from all possible tag sequences t_1^n associated with the word sequence to analyze w_1^n according to a dictionary, the most probable one

$$\hat{t}_1^n = \arg\max_{t_1^n} P[w_1^n, t_1^n] \quad (5)$$

$$= \arg\max_{t_1^n} \prod_{i=1}^{n} P[w_i|t_i] P[t_i|t_{i-N+1}^{i-1}] \ . \quad (6)$$

Building such a tagger is thus done from two models estimating $P[w_i|t_i]$ and $P[t_i|t_{i-N+1}^{i-1}]$. We chose the parameters of these models by optimizing them according to tag accuracy measured on a 40-minute development corpus. $P[w_i|t_i]$ are computed from counts in a tagged training corpus, by using additive smoothing, while $P[t_i|t_{i-N+1}^{i-1}]$ are computed from 3-gram LMs, by using the original Kneser-Ney smoothing method (Chen and Goodman, 1998).

In order to adapt the tagger to the characteristics of spoken documents, we used a 200,000-word training set from the manual transcriptions of the training part of the ESTER corpus. Moreover, we removed all capital letters and punctuation marks to obtain a format similar to a transcription and segment the text into breath-groups. We also restrained

transcription	manual	automatic
HMM tagger	95.7 (95.9)	95.7 (95.9)
simple tagger	90.6 (91.0)	90.7 (91.1)
Cordial	90.7 (95.0)	90.6 (95.2)

Table 1: Tag accuracy (%), where results between parentheses are computed when confusion between common names and proper names are ignored.

the vocabulary of the tagger to the one of our ASR system. We chose our POS tags in order to distinguish the gender and the number of adjectives and nouns, and the tense and the mood of verbs, which led to a set of 93 tags.

4.2. Evaluation

We quantitatively evaluate morphosyntactic tagging on a 1-hour show, available in two versions: one manually transcribed by a human annotator and one generated by our ASR system with a word error rate (WER) of 22.0 %. To measure tag accuracy, we manually tagged the version produced by the annotator. We first investigated the behavior of the tagger on the manually transcribed text by comparing the tag found for each word with the one of the reference. For the automatic transcription, evaluating the tagger is more problematic than for the manual transcription since the ASR output contains misrecognized words; for the ungrammatical output hypotheses, it becomes impossible to know which POS sequence would be right. We therefore compute the tag rate only for the words that are correctly recognized.

Results obtained over the 1-hour test corpus (Tab. 1, first line) show a tag accuracy over 95 % which is comparable to the numbers usually reported on written corpora. Furthermore, we get the same results on both manually and automatic transcriptions, which establishes therefore that morphosyntactic tagging is reliable, even for text produced by an ASR system whose recognition errors are likely to jeopardize the tagging of correctly recognized words. The robustness of tagging is explained by the fact that tags are locally assigned. One reason that might explain the identical results obtained for both transcriptions is that the one produced by human annotator contains out-of-vocabulary words; this does not occur for the automatic transcriptions where the set of possible words is limited to the vocabulary of the ASR system that is also the one of the tagger. However, the number of words missing in the lexicon is quite low for the studied corpus and is responsible for only 52 of the 481 tagging errors.

To measure how difficult is the tagging task from the available tagged lexicon and training corpus, we compared the results previously obtained with a simple approach which associates each word with its most frequent tag according to the training corpus. The tag accuracies measured with this simple method already exhibits good results over 90 % (Tab. 1, line 2). Nevertheless, they also show that the use of HMM makes tagging errors decrease by more than 50 %. Furthermore, we compared our tagger with Cordial[1], one of the best taggers available for written French which has already given good results on a spoken corpus (Valli and Véronis, 1999). The last line of Table 1 shows results comparable with our HMM-based tagger when we ignore confusion between proper names and common names. Indeed, the lack of capital letters is particularly problematic for Cordial, which relies on this information to detect proper names.

[1]Distributed by *Synapse Développement* corporation.

transcription	manual	automatic
original tag set	95.7	95.7
number errors ignored	96.1	96.1
gender errors ignored	96.3	96.4
conjugation errors ignored	96.1	96.0
number and gender errors ignored	96.7	96.8

Table 2: Different tag accuracies (%) measured with our HMM tagger.

Generally, tag accuracy mainly depends on the chosen tag set. To evaluate how the granularity of the tag set acts upon tagging performance, we give in Table 2 the results observed by ignoring some errors. The last line can for instance be interpreted as the tag accuracy computed from a tag set without any information about gender and number; results obtained with this last tag set are close to 97 % and are better than the original ones. Nevertheless, as the information about gender and number is relevant to improve transcriptions, we keep this information. We extend instead the 93-tag set by adding specific classes for the 100 most frequent words. Firstly, this reduces the ambiguities of grammatical words according to their POS, while some of them are very difficult to disambiguate; we measure on the automatic transcription a 97.1 % of correct tags with this extended tag set, instead of the 95.9 % previously given with the original set. Secondly, this new tag set allows us to introduce explicit knowledge from the POS sequences about syntax through the use of some common grammatical words.

5. Improvement of transcriptions

After showing that transcription hypotheses produced by ASR systems can be reliably tagged, we exhibit in this section the interest of POS knowledge to improve ASR outputs. To do this, we use morphosyntactic information in a post-processing stage of an ASR system to rerank the $\mathcal{N}$-best word sequences generated for each breath-group. Each entry of such $\mathcal{N}$-best lists can be seen as a standard text, enabling thus POS tagging. We describe in this section how we resort to morphosyntax to select better hypotheses and demonstrate experimentally that this information is relevant to improve transcriptions. We finally discuss the changes induced by POS over recognized words.

5.1. Method

To exploit morphosyntactic resources for each breath-group, we first determine the most likely POS tag sequence t_1^m corresponding to a word sequence w_1^n. Based on this

information, we compute the morphosyntactic probability

$$P[t_1^m] = \prod_{i=1}^{m} P[t_i | t_{i-N+1}^{i-1}] \ . \quad (7)$$

To take into account longer dependencies than the 4-gram word-based LM used by our ASR system, we chose a 7-gram POS-based LM.

We propose an original score of a hypothesis w_1^n (Huet et al., 2007) by adding the morphosyntactic score with an appropriate weight to the LM and acoustic scores:

$$\begin{aligned} s(w_1^n) &= \log P(y_1^t | w_1^n) + \alpha \log P[w_1^n] \\ &\quad + \beta \log P[t_1^m] + \gamma n \ . \end{aligned} \quad (8)$$

Contrary to previous approaches, POS information is here introduced at the breath-group level and not at the word level. This property allows us to more explicitly penalize unlikely sequences of tags like a plural noun following a singular adjective, and to differently tokenize sequences of words and tags by associating a unique POS with locutions, consecutive proper names or cardinals.

We also propose another score, by using lexical probabilities $P[w_1^n | t_1^n]$ which are usually included in class-based LM (defined by Eq. (3)):

$$\begin{aligned} s'(w_1^n) &= \log P(y_1^t | w_1^n) + \alpha \log P[w_1^n] \\ &\quad + \beta (\log P[t_1^m] + \log P[w_1^n | t_1^n]) + \gamma n \ . \end{aligned} \quad (9)$$

This second score becomes close to the linear interpolation of log-linear probabilities of an acoustic model, a word-based language model and a class-based language model.

Based on the score function defined in (8) or (9), we can reorder $\mathcal{N}$-best lists using various criteria. We consider three criteria commonly used in ASR, namely maximum a posteriori (MAP), minimum expected word error rate (Stolcke et al., 1997) and consensus decoding on $\mathcal{N}$-best lists (Mangu et al., 2000). The first criterion selects among the $\mathcal{N}$-best list generated for each breath-group the best hypothesis $w^{(i)}$ which maximizes $s(w^{(i)})$ or $s'(w^{(i)})$. The two last ones, often used in current systems, aim at reducing the word error rate to the expense of an increased sentence error rate. They both estimate the number of misrecognized words by aligning several hypotheses generated by the ASR system for each given breath-group, with the objective to reduce the WER rather than to maximize a log probability score.

5.2. Experiments

To test our method, we used the IRENE broadcast news transcription system, jointly developed by IRISA and Telecom for the ESTER evaluation campaign. We use in our experiments a 4-gram word-based language model built by interpolating a LM estimated on 1 million words from the manual transcriptions of the training set of the ESTER corpus with a LM estimated from 350 million words from the French newspaper *Le Monde*. The LM used to compute the morphosyntactic probability $P(t_1^m)$ was built from a 200,000-word extract from the training set of the ESTER corpus. Furthermore, the parameters α, β and γ of (8) or (9) were estimated on a development set of 4 hours, while the test data consist of 10 hours, two of which were produced by different broadcasters from the development data. Let us note that the test data were produced one year later than the training and development corpora. A more complete description of the decoding process of our ASR system can be found in Huet et al. (2007). All the experiments presented here were done from 1,000-best lists.

baseline ASR system	24.7
$s(w_1^n)$	24.0
$s'(w_1^n)$	23.9
class-based LM	24.2

Table 3: WER(%) on test data by reordering 1000-best lists with a MAP criterion.

We reordered the $\mathcal{N}$-best lists generated for the test corpus according to the MAP criterion, with a score ignoring POS (Tab. 3, line 1), or by taking into account this information without $\log P(w_1^n | t_1^n)$ (Tab. 3, line 2) or with $\log P(w_1^n | t_1^n)$ (Tab. 3, line 3). The results show that our approach reduces the number of misrecognized words. To assess this decrease, we carried out statistical tests, assuming independence of the errors across breath-groups. Both the paired t-test and the paired Wilcoxon test indicate with a p-value less than 0.1 % that the improvement of the WER due to morphosyntactic resources is significant.

Besides, we compared our approach with the method relying on a class-based LM (Maltese and Mancini, 1992), traditionally used to include POS in the ASR process (Tab. 3, last line). This class-based LM was built from the same 200,000-word corpus as the LM computing $P[t_1^m]$ and with the same tag set; it is here used by linearly interpolating it with a word-based LM according to

$$\begin{aligned} P[w_1^n] &= \prod_{i=1}^{n} [\lambda P_{\text{word}}[w_i | w_1^{i-1}] \\ &\quad + (1-\lambda) P_{\text{class-based}}[w_i | w_1^{i-1}]] \end{aligned} \quad (10)$$

where P_{word} is given by the 4-gram LM and $P_{\text{class-based}}$ is obtained by (3) with $N = 7$. This method leads to a lower decrease of the WER than our approach. The paired t-test and the paired Wilcoxon test indicate with respective p-values of 0.2 % and 1.2% that $s'(w_1^n)$ is statistically better than the score using the class-based LM. The same experiments led with $s(w_1^n)$ are less clear since the paired t-test and the paired Wilcoxon test respectively give p-values of 4.5 % and 8.6 %.

Table 4 reports results obtained by our approach from 1,000-best lists for the three decoding criteria. Interestingly, results show that including morphosyntactic information improves the quality of transcriptions, whatever the decoding criterion used. Statistical tests were carried out, assuming independence of the errors across breath-groups, and showed that the difference of WER by using or not POS knowledge is statistically significant for all the tested decoding criteria with p-values lower than 0.1 %.

	WER			SER		
	MAP dec.	min. WE	cons. dec.	MAP dec.	min. WE	cons. dec.
without POS	24.7	24.2	24.0	70.5	70.4	71.1
with POS	23.9	23.5	23.5	69.1	69.6	70.2

Table 4: WER (%) and sentence error rates (%) on test data for various decoding criteria.

5.3. Analysis of the results

We study in this section the changes induced by the morphosyntactic resources with scores $s(w_1^n)$ and $s'(w_1^n)$. As previously shown, this information yields a significant decrease of the WER. This improvement tends to translate into the production of more grammatical utterance transcriptions as indicated by the sentence error rates reported in Table 4. Indeed, an analysis of the sentence error rate shows a significant reduction when taking into account morphosyntactic information. A manual analysis confirms this trend. We give in Figure 2 examples of breath-group transcriptions modified by POS knowledge, by showing successively the reference transcription, the output of the ASR system generated without taking into account POS information and the output generated by including POS knowledge. The two first examples exhibit a correction of an agreement error over the noun "*minorités*" and the adjective "*polynésien*". The third shows a correction of a confusion between infinitive and past participle for "*annoncer*". In the last example, the use of the preposition "*dans*" is here rectified by POS knowledge.

To measure quantitatively the type of errors corrected by morphosyntactic resources, we decided to ignore confusion about inflections. This leads us to define two new metrics. The first one, called *lemma error rate* (LER), is defined similarly to the WER by still computing the number of insertions, deletions or substitutions, but on lemmas rather than on words. To do this, we tag the reference and automatic transcriptions with our morphosyntactic tagger and resort to the FLEMM lemmatizer, which is commonly used for the French language (Namer, 2000), to lemmatize the reference transcription and the one produced by the ASR system. Since these natural language processing techniques are automatic, errors are committed, which slightly disturbs the LER computation. However, our tagger performs well on broadcast news, as we previously showed, and numerous tagging errors affect grammatical words that are easy to lemmatize. The main interest of the LER is that it ignores numerous agreement or conjugation errors. The second metrics, called LER_{lex}, specifically measures the errors on words that mainly bring lexical meaning to the document to transcribe; it is computed from the LER by limiting the reference and the output of the ASR system to nouns, verbs and adjectives. Auxiliaries and modal verbs are also discarded as they are function words.

Results obtained on the test corpus according to these two new metrics are given columns 2 and 3 of Table 5. The comparison between the WER and the LER shows that for the baseline ASR system, 2.9 % of words (from 24.7 % to 21.8 %) are correct according to their lemma, but have a wrong inflection. This figure is reduced to 2.6 % by using morphosyntax, which indicates that this information corrects some agreement errors. The use of morphosyntax leads also to an absolute decrease of the WER by 0.7 % or 0.8 % according to the score used, which respectively translates into an absolute decrease of the LER by 0.4 % or 0.5 %. The comparison of these values suggests that globally around 40 % of the improvement due to morphosyntax concerns errors about inflections. Interestingly, a comparison of LER_{lex} shows a different influence of morphosyntactic information according to the score used; indeed, $s'(w_1^n)$ leads to a much more important decrease than $s(w_1^n)$. This indicates that $s'(w_1^n)$ tends to modify the lemmas of lexicalized words, while $s(w_1^n)$ acts more upon grammatical words.

	WER	LER	LER_{lex}
baseline	24.7	21.8	22.9
$s(w_1^n)$	24.0	21.4	22.6
$s'(w_1^n)$	23.9	21.3	21.8

Table 5: WER (%), LER (%) and LER_{lex} measured on test data by reordering 1,000-best lists with a MAP criterion.

6. Conclusion

In this paper, we have established that the language resource of POS improves the output of an ASR system for the French language. We have first shown that transcriptions, despite their oral characteristics, can be reliably tagged by morphosyntactic information. This property means that NLP methods based on tagged texts can be used on spoken documents. We have studied here a particular use of POS tagger and we have demonstrated that morphosyntactic information leads to improve ASR results by generating more grammatical hypotheses. In additional experiments (Huet et al., 2007) not reported here, we also noticed that POS tagging, similarly used in a post-processing stage of ASR, helps the confidence measure computation. This measure, which indicates how reliable is a transcribed word, is an interesting clue for NLP processing of automatic transcriptions.

An important restriction of our method is that we use morphosyntactic resources by considering independently each breath-group. However, some errors are produced because dependencies across breath-groups are not taken into account. A way to resolve this problem would be to consider a more linguistic segmentation than the one based on the silence duration alone.

7. References

T. Brants. 2000. TnT - a statistical part-of-speech tagger. In *Proc. of the Conference on Applied Natural Language Processing (ANLP)*.

```
REF    : à part quelques MINORITÉS
w/o POS: à part quelques MINORITÉ
w   POS: à part quelques minorités
REF    : aujourd'hui le vieux lion POLYNÉSIEN  semble des PLUS RÉVEILLÉ
w/o POS: aujourd'hui le vieux lion POLYNÉSIENS semble des **** PYRÉNÉES
w   POS: aujourd'hui le vieux lion polynésien  semble des **** PYRÉNÉES
REF    : IL ME reste quelques secondes pour vous ANNONCER
w/o POS: ** LE reste quelques secondes pour vous ANNONCÉ
w   POS: ** LE reste quelques secondes pour vous annoncer
REF    : DANS un état
w/o POS: ANS  un état
w   POS: dans un état
```

Figure 2: Examples of transcription errors corrected with morphosyntactic resources.

P.F. Brown, V.J. Della Pietra, P.V. deSouza, J.C. Lai, and R.L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

S.F. Chen and J. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical report, Harvard University, Cambridge, MA, USA.

S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. 2005. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In *Proc. of Interspeech*.

J. Giménez and L. Màrquez. 2004. SVMTool: A general POS tagger generator based on support vector machines. In *Proc. of LREC*.

S. Huet, G. Gravier, and P. Sébillot. 2007. Morphosyntactic processing of N-best lists for improved recognition and confidence measure computation. In *Proc. of Interspeech*.

G. Maltese and F. Mancini. 1992. An automatic technique to include grammatical and morphological information in a trigram-based statistical language model. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

L. Mangu, E. Brill, and A. Stolcke. 2000. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400.

B. Merialdo. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–171.

F. Namer. 2000. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues*, 41(2):523–547.

A. Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

A. Stolcke, Y. König, and M. Weintraub. 1997. Explicit word error minimization in N-best list rescoring. In *Proc. of Eurospeech*.

A. Valli and J. Véronis. 1999. Étiquetage grammatical de corpus oraux : problèmes et perspectives. *Revue française de linguistique appliquée*, 4(2):113–133.